A Review: Data Mining Brainwave

Patricia Saibul Cyril, Abdul Samad Shibghatullah and Norhaziah Md Salleh
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Karung Berkunci 1200, Hang Tuah Jaya, Ayer Keroh, 75450, Melaka
paty26o5@yahoo.com
{samad|haziah}@utem.edu.my

***Abstract*— Data mining has been long a topic of many international conferences. The term data mining has pressure many researchers to take part in discovering knowledge. Since data mining become an active research area, many tools have been introduced to the worldwide. However, in order to use such a data mining tool, we must first understand the concepts of data mining, how data mining works and what techniques does it provides. Fundamentally, data mining is related with data, information and knowledge, which concerns on the process of finding relationships and discovering hidden patterns that exists in a large database.**

This paper presents the definition of data mining from many points of view and covers only six techniques that have common application in the real world situation. In addition, some of data mining application will be cover at the end of this paper.

***Index Terms*— Association rules; clustering; data; data mining; decision tree;**

I. INTRODUCTION

The world is going fast. The faster the world goes, the more data produced, collected and warehoused. We believe that the modern world is a data-driven one since we are surrounded by data, numerical and otherwise, which must be analyzed and processed to convert it into information that informs, instructs, answer, or otherwise aids understanding and decision-making [8].

At the current stage, lack of data is no longer a problem [1]. We got warehouse to store collected data, and however we have no knowledge to discover useful information from the data. Hence, people were desperate to develop new technologies and tools in order to process data into useful information and knowledge automatically. Although data mining is a relatively new term, the technology is not. Data mining [2] is a combination of database and artificial intelligence technologies.

Generally, data mining can be referred as the process of searching and extracting valuable information from a large amount of data by performing tasks and techniques. In the other hand, It is also refers as a process of extracting information and discovering hidden [1]patterns and relationships in large databases in order to make better and more informed decisions [14].

In order to create a successful accomplishment in formative and solving problems related with data, data mining manage to have many techniques that will be used to automatically extract the knowledge from data stored in the database. In this paper, data mining represents five techniques that are used to enhance the performance in searching and discovering hidden patterns as well as in decision-making.

II. DATA MINING DEFINITION

Many researchers describe data mining in different approaches and opinions. Since decades, more than 50 researchers participate in revising the topic of data mining. From the studies, researchers have defined data mining through their understanding. In this section, we will look on a few data mining definitions from six researchers.

A. Seifert [4] defines data mining in which it involves the use of sophisticated data analysis

ISBN 978-967-5048-36-4

tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms and machine learning methods.

B. *Defit and Sap* [6] define data mining as searching for valuable information in large volumes of data. It is the process of extraction of implicit and potentially useful information such as knowledge, rules, constraints and regularities from data stored in repositories.

C. *Roiger and Geatz* [12] define data mining as the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data contained within a database.

D. *Kantardzic* [8] defines data mining as a process of discovering various models, summaries and derived values from a given collection of data.

E. *Edelstein* [16] defines data mining as a process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that are used to understand what customer want and predict what they will do.

F. *Guidici* [11] defines data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the databases.

III. WHY DATA MINING?

Data mining is not a mysterious manner. The way it performs almost same as the human being does. In recent year, data mining has attracted a great deal of attention in the information industry due to the wide availability of huge data amounts and the imminent need for turning such data into useful information and knowledge [3]. It is the process of learning from the past experiences and applying the knowledge to another situation. The knowledge gained will be used either in decision-making or problem solving, or both.

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the [2]database industry in the development of the following functionalities: data collection and data creation, data management, and data analysis and understanding [3].

In the corporate world, there are a large amounts of data captured in enterprise databases. These databases are too large for traditional statistical techniques. Data mining for sure can identify the patterns in the data for certain targets such as profitable or unprofitable.

In institutional research, it involved with a large numbers of variables. Our constraints are we have insufficient time and resources to investigate all the relationships that might be informative [15].

IV. DATA MINING TASKS AND TECHNIQUES

In this section, we review and discuss the most common data mining tasks such as association rules and clustering and techniques such as genetic algorithm, decision tree, neural networks and statistical technique.

A. Association rules

Association rules were first introduced as a means of determining relationships among a set of items in a database [9]. Like clustering, association rules are a form of unsupervised learning. Association rules can have one or several output attributes. The output attribute for one rule can be an input attribute for another rule [12].

One of the analysis which make association rules as a popular technique is the market basket analysis. A market analysis is a collection of items purchased by a customer in a single transaction, which is a well-defined business activity [8]. For example, [1] if a customer buys

Patricia Saibul Cyril, Dr. Abdul Samad Shibghatullah, and Associate Professor Norhaziah Md Salleh are with the Department of Information and Communication Technology, University of Technical Malaysia Melaka (UTeM), 75450 Ayer Keroh, Melaka, Malaysia. (e-mail: paty26o5@yahoo.com, samad@utem.edu.my; haziah@utem.edu.my).

one brand of beer, he/ she usually buys another brand of chips in the same transaction.

TABLE 1
SAMPLE DATABASE OF PURCHASE TRANSACTIONS

Transaction	Items
t_1	Beers, Chips
t_2	Beers, Chips, Peanuts
t_3	Chips, Coke, Pretzels
t_4	Coke, Pretzels

Formally, give a set of m items $I = \{I_1, I_2, ..., I_m\}$ and a database of n transactions $D = \{t_1, t_2, ..., t_3\}$, where a given transaction contains k items $t_i = \{I_{i1}, I_{i2}, ..., I_{ik}\}$ and $I_{ij} \in I$, an association rule is an implication of the form $X \Rightarrow Y$ where $X, Y \subset I$ are sets of items called *itemsets* and $X \cap Y = \phi$ [9].

The *support, s* of an item or set of items represents the percentage of transactions in which the item can be found. For example, {Chips} has 75 percent supports when it appears in almost all but the last transaction {Peanuts, Pretzels} has 0 percent of support since the two never appear in the same transaction.

Association rules are inferred based on support and confidence [9]. Support for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$ [9]. Confidence α refers to the strength of co-occurrence between two items or sets of items. The confidence for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain [9].

We can use probabilistic notation to represent the support and confidence. [9] The support of itemset X can be written as

$$P(X) = \frac{\text{number of transactions in which } X \text{ occurs}}{\text{Number of transactions}}$$

Now, we calculate the confidence of a rule $X \Rightarrow Y$ between Beers and Chips. Chips occur in three transactions, therefore the support of item Chips is 75%.

$$P(X|Y) = \frac{X \cup Y}{P(Y)}$$

$$P(\text{Chips}|\text{Beers}) = \frac{P(\{\text{Beers,Chips}\})}{P(\text{Chips})}$$

$$= \frac{2}{3} = 67\%$$

B. Clustering

Clustering is the unsupervised classification of patterns into groups based on upon similarity, where a pattern is a representation of features or observations made on an object [13]. Clustering is used in several exploratory data analysis tasks, customer retention and management, and web mining [13].

Clustering is the process of creating a partition so that all the members of each set of the partition are similar according to some metric. Samples for clustering are represented as a vector of measurements, or more formally, as a point in a multidimensional space. Samples within a valid cluster are more similar to each other than they are to a sample belonging to a different cluster [8].

A simple example of clustering information for nine customers, distributed across three clusters is showed in table 2 [8].

TABLE 2
SAMPLE SET OF CLUSTERS CONSISTING OF SIMILAR METRIC

	# of Items	Price
Cluster 1	2	1700
Purchase few high-priced	3	2000
items	4	2300
Cluster 2	10	1800
Purchase many high-priced	12	2100
items	11	2500
Cluster 3	2	100
Purchase few low-priced	3	200
items	3	350

Clustering approaches may be broadly categorized into two methods: hierarchical and partitional [9].

1) Hierarchical Algorithms: Hierarchical clustering algorithms create nested sets of clusters, producing a binary tree structure known as a *dendrogram* [9]. In the dendrogram, each node represents a cluster.

2) Partitional Algorithms: Partitional clustering algorithms produce a single partition of the patterns [9]. Unlike hierarchical clustering, partitional clustering requires less time and space when working on large data sets.

C. Genetic Algorithm

Genetic algorithms apply an evolutionary approach to inductive learning [12]. It was introduced as a computational analogy of adaptive systems. This genetic algorithm learning is based on the Darwinian principle of natural selection.

One population of rules is initially created randomly to represent a possible solution to a problem. To produce offspring for the next generation, there will be combinations between pairs of rules. The pairs of rules are usually the strongest rules selected as parents [1]. Genetic algorithm learning are modeled loosely on the principles of Darwinian, that employing a population of individuals that undergo selection in the presence of variation-inducing operators such as mutation and recombination (crossover) [7].

A mutation process is used to randomly modify the genetic structures of some members of each new generation [1]. The following is the process in genetic algorithms [7]:

1. Randomly generate an initial population M(0)
2. Compute and save the fitness u(m) for each individual m in the current population M(t)
3. Define selection probabilities p(m) for each individual m in M(t) so that p(m) is proportional to u(m)
4. Generate M(t+1) by probabilistically selecting individuals from M(t) to produce offspring via genetic operators
5. Repeat step 2 until satisfying solution is obtained.

Genetic algorithms are appropriate for problems that require optimization with respect to some computable criterion [1]. For example, to obtain appropriate solutions in a reasonable amount of time, large and complex problems require a fast computer [1]. Due to the availability of affordable high-speed computers, genetic algorithms have been applied in mining large data sets recently.

In addition, by using these conventional techniques, genetic programming has successfully applied to problems that are difficult to solve. Common areas of application include scheduling problems, such as the traveling salesperson problem, network routing problems for circuit-switched networks and problems in the area of financial marketing [12].

D. Decision Tree

Decision tree is a flow-chart-like tree structure used for classification, clustering, feature selection, and prediction [9]. It is a simple knowledge representation, also known as predictive model that is, a mapping from observations about items to conclusions about its target value. It depicts rules for dividing training data into groups based on the regularities in the data [9]. Two ways using decision tree: categorical response variables and continuous response variables. When the response variables are continuous, the decision tree is often referred to as a *regression tree* [9]. If the response variables are categorical, it is called a *classification tree* [9]. Both types of trees however, applied the same concepts of decision tree.

A decision tree consists of internal nodes which denote a test on an attribute, branch which represents an outcome of the test and internal nodes that represent class labels or class distribution. The root and the internal nodes are labeled with questions in order to find a solution to the problem under consideration [9].

Fig. 1 shows a simple decision tree for classification of samples with two input attributes X and Y [8]. All samples with feature values X>1 and Y=B belong to class2, while the samples with values X<1 belong to Class1, whatever the value for feature Y [8].

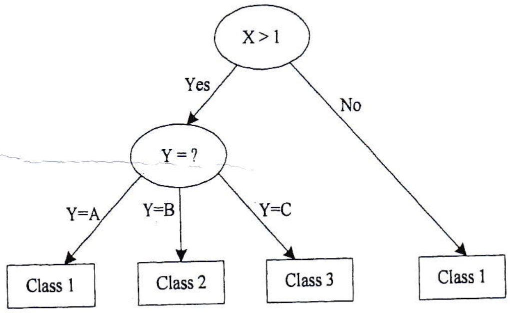


Fig. 1. A simple decision tree with the tests on attributes X and Y

E. Neural Networks

Neural networks are an approach to computing that involves developing mathematical structures with the ability to learn. Neural networks offer a mathematical model that attempts to mimic the human brain [12]. The brain is a highly complex, nonlinear, and parallel information-processing system. It has the capability to organize its components so as to perform certain computations with higher quality and many times faster that the fastest computer in existence today [8].

Neural networks have been successfully applied to problems across several disciplines and for this reason are quite popular in the data mining community [12]. In the other hand, neural networks can be constructed for either supervised learning or unsupervised learning, and also come in many shapes and forms. Fig. 2 [12] shows a fully connected feed-forward neural network of three layers

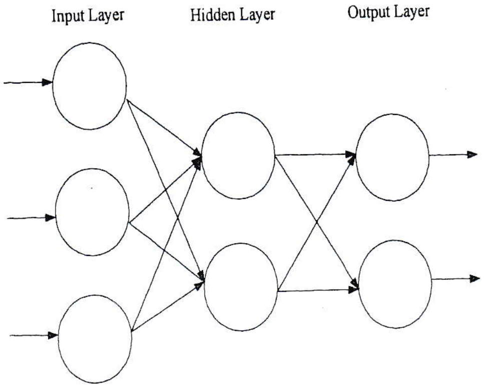
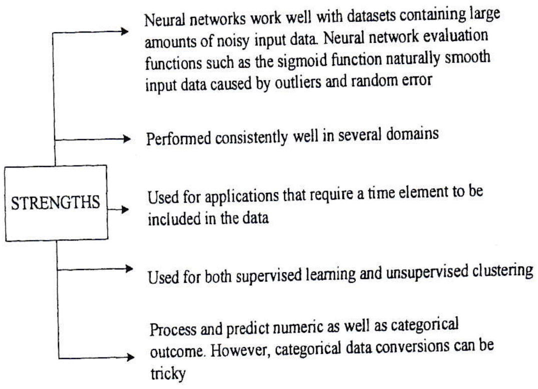


Fig. 2 A Multilayer fully connected neural network

The nodes at one layer are connected to all nodes at the next layer. For this reason, the network is known as fully connected. In addition, each network node connection has an associated weight [12]. However, nodes within the same layer of the network in fig. 3 are not connected to one another.

There are two phases that operating in the neural network. The first phase is called the learning phase [12]. The input values associated with each instance enter the network at the input layer during network learning. For each input attribute that contained in the data, there is one layer node exists for it. The neural network uses the input values together with the network connection weights to compute the output for each instance and the output for each instance is compared with the desired network input [12].

Reference [12] brief lists of the strengths and weaknesses for this technique as the following in fig.3 and fig.4.



Neural networks work well with datasets containing large amounts of noisy input data. Neural network evaluation functions such as the sigmoid function naturally smooth input data caused by outliers and random error

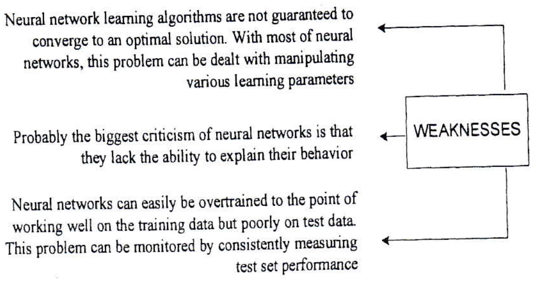
Performed consistently well in several domains

Used for applications that require a time element to be included in the data

Used for both supervised learning and unsupervised clustering

Process and predict numeric as well as categorical outcome. However, categorical data conversions can be tricky

Fig. 3 Strengths of Neural Networks Techniques



Neural network learning algorithms are not guaranteed to converge to an optimal solution. With most of neural networks, this problem can be dealt with manipulating various learning parameters

Probably the biggest criticism of neural networks is that they lack the ability to explain their behavior

Neural networks can easily be overtrained to the point of working well on the training data but poorly on test data. This problem can be monitored by consistently measuring test set performance

Fig. 4 Weaknesses of Neural Networks Techniques

F. Statistical Techniques

According to Kantardzic [8], statistics is the science of collecting and organizing data and drawing conclusions from data sets. Statistics is an indispensable component in data selection, sampling, data mining and extracted knowledge evaluation [1].

In statistics, the term sample refers to a subset of a population and it is used to describe a finite data set of n-dimensional vectors. Sometimes, sample is called as data set. From the given sample, statistical model of the population is build to help in making inferences concerning on the same population. If our inferences from the data set are to be valid, we must obtain samples that are representative of the population [8].

Statistic inference is the main form of reasoning relevant to data analysis [8]. The theory of statistical inference consists of those methods that can be categorized into two major areas which known as estimation and tests of hypotheses. Those methods make inferences or generalizations about a population.

1) Estimation: The goal of estimation is to gain information from a data set T in order to estimate one or more parameters w belonging to the model of the real world system f(X, w) [8]. With estimation techniques, statistics can also deal with problem such as missing data.

2) Statistical Testing: In statistical testing, one has to decide whether a hypothesis concerning the value of the population characteristic should be accepted or rejected in the light of an analysis of the data set [8]. The structure of hypothesis is formulated with the use of the term *null hypothesis,* H_0 which refers to any hypothesis to test. If the given data set contains strong evidence which proving the hypothesis is not true, and then H_0 will be rejected. The rejection of H_0 leads to the acceptance of an alternative hypothesis which denoted by H_1 about the population [8].

V. DATA MINING APPLICATIONS

Currently, many businesses and scientific communities are employing data mining technology. Their number continues to grow, as more and more data mining success stories become known [8]. In this section, a few application domains will be examined and illustrated by the implemented data mining system's result.

A. Financial Data Analysis

Mellon Bank has used the data on existing credit-card customers to characterize their behavior and they try to predict what they will do next. Using IBM Intelligent Miner, Mellon developed a credit card-attrition model to predict which customers will stop using Mellon's credit card in the next few months. Based on the prediction results, the bank can take marketing actions to retain these customers' loyalty [8].

B. Telecommunications Industry

Worldcom is another company that has found great value in data mining. By mining databases of its customer-service and telemarketing data, Worldcom has discovered new ways to sell voice and data services. For example, it has found that people who buy two or more services likely to be relatively loyal customers. It also found that people were willing to buy packages of products such as long-distance, cellular-phone, Internet, and other services. Consequently, Worldcom started to offer more such packages [8].

C. Scientific Applications

The Gamma ray bursts are brief gamma ray flashes that originate outside of our solar system. More than 1000 such events have been recorded. A widely held belief in the scientific community was that there were two classes of gamma ray bursts. The third gamma ray burst class been discovered by using statistical cluster analysis [12].

D. Sports and Gaming

The gaming industry has incorporated historical models of customer gambling trends to determine how much an individual customer should be spending while visiting their favorite casino [12].

E. Fraud Detection

The Aspect (Aspect Security for Personal Communications) European research group has employed unsupervised clustering to detect fraud in mobile phone networks. For each user, the system stores a user history as well as a usage profile. Fraudulent behavior is suspected with marked differences between current usage and user history [12].

VI. CONCLUSION

Data mining is very useful in solving real world problems, other than just in the commercial world where gaining competitive advantages is crucial for good performance. Data mining techniques involve two types of learning, supervised and unsupervised learning that been applied either in association rules, clustering, genetic algorithm, decision tree, and statistical techniques, or may be both for neural networks techniques. As far as data mining continue to exist, we know that data mining has been successfully applied to problems across several regulations. For this reason, we believe that data mining is still a very long way to discover and for that, there must be no excuses for us no to have any ideas about data mining.

REFERENCES

[1] J. L. Sang and S. Keng, (2001). "A Review of Data Mining Techniques", University of Nebraska-Lincoln, Nebraska, USA. [Online] Available: http://www.emerald-library.com/ft

[2] "Data Mining". [Online] Available: http://www-pub.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/10.htm

[3] M. Hanna (2004), "Data Mining in the E-Learning Domain", Vol.21. p. 29-34 [Online] Available: http://emeraldinsight.com/1065-0741.htm

[4] J, W. Seifert, "CRS Report for Congress".2003, Retrieved from the Library of Congress.

[5] R. Connelly, "Introduction Data Mining", Department of Mathematics and Computer Science, Providence College, vol.19. p. 87-96, 2004 [Online] Available: http://portal.acm.org/citation.cfm?id=1060095

[6] S. Defit and M. N. Sap, "Data Mining: A Preview", University of Technology Malaysia, Malaysia, vol. 12, no. 1, 2000.

[7] T. Wong and H. Wong, "Application of Genetic Algorithm", vol. 4, [Online] Available: http://www doc.ic.ac.uk/~nd/suprise_96/journal/vol4/tcw2/report.html#Overview

[8] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms.* U.S: Wiley-Interscience, 2003, p. xi.

[9] M. W. Berry and M. Browne, *Lectures Notes in Data Mining.* Singapore: World Scientific Publishing Co. Pte. Ltd., 2006.

[10] J. Hartigan, *Clustering Algorithms*, New York: John Wiley & Sons, 1975

[11] P. Giudici, *Applied Data Mining: Statistical Methods for Business and Industry,* John Wiley& Sons Ltd, West Sussex, England, 2003

[12] R. J. Roiger and M. W. Geatz, *Data Mining: A Tutorial-based Primer.* U.S: Addison-Wesley, 2003, p. 4.

[13] S. K. Pal and P. Mitra, *Pattern Recognition Algorithms for Data Mining,* U.S: Chapman and Hall/CRC, 2004.

[14] B. W. Yap, "Some Applications of Data Mining", presented at the National Statistics Conference, Putrajaya International Convention Centre, Putrajaya, Malaysia, 2006.

[15] E. Thomas, "Data Mining: Definitions and Decision Tree Examples" [Online] Available: http://airpo.binghamton. edu/conference/jan2004/Thomas_data_mining.pdf

[16] H. Edelstein, "Building profitable customer relationships with data mining", white paper-executive briefing, SPSS, 2000 [Online] available at: www.spss.fi/PDF/Building proftable_cust_relations_DM.pdf

BIOGRAPHIES

PATRICIA SAIBUL CYRIL is a postgraduate student at Faculty of Information and Communication Technology, UTeM. She graduated from the Faculty of Information and Communication Technology, and studied at the University of Technical Melaka Malaysia (UTeM). Currently, she is pursuing her study in master's degree from the same university, UTeM.

Her employment experience included the Sabah Advancement of Information Technology (IT) Unit Organization, known as KIT under the Government of Sabah. Her special fields of interests included social services, science politic, database and mathematic.

DR. ABDUL SAMAD SHIBGHATULLAH is a lecturer at Universiti Teknikal Malaysia Melaka. He performed research and teaching in the Faculty of Information and Communication

Technology. He received a B. Accounting from Universiti Kebangsaan Malaysia of Bangi. He received his M.Sc. in Computer Sciences from Universiti Teknologi Malaysia, Skudai, and received his Doctor of Philosophy at Brunel University. His research interests include scheduling, data mining, transportation and optimization.

NORHAZIAH MD SALLEH is an associate professor at Universiti Teknikal Malaysia Melaka. Formerly, she is a Chief Analyst at Universiti Utara Malaysia, Kedah and currently, teaching database subjects at Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka.

She received a B.Sc. in Computer Science from Indiana State, United States and a M.Sc. in Computer Science from University of Bradford, United Kingdom. Her special fields of interests included database, data mining, information hiding, Software Engineering, Systems Analysis and Design, Life Assurance, and ISO - Internal Auditing.

ISBN 978-967-5048-36-4